Математическая биология и биоинформатика
2017. Т. 12. № 2. С. 411–434. doi: /10.17537/2017.12.411

==========**ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ** ==========

УДК: 004.02, 004.94

Сравнительный анализ подходов к классификации дифракционных изображений биологических частиц, получаемых в экспериментах по когерентной рентгеновской дифракционной микроскопии

Бобков С.А.*

Национальный исследовательский центр "Курчатовский Институт", Москва, Россия

Аннотация. Метод когерентной рентгеновской дифракционной микроскопии дает возможность определения трехмерной структуры наноразмерных объектов, в том числе таких биологических частиц, как белки и вирусы, с разрешением до 1 Å. В таких экспериментах излучение лазера на свободных электронах рассеивается на изучаемых объектах. На основе собранных дифракционных изображений можно определить исходную структуру объекта. Однако далеко не все дифракционные изображения, получаемые в экспериментах, подходят для восстановления структуры. Большинство изображений бывают пустыми, многие изображения относится к частицам примеси, другие содержат дифракционную картину от нескольких частиц. Таким образом, классификация изображений по типу структуры становится важным этапом первичной обработки данных. В работе сравниваются несколько подходов к классификации изображений по типу структуры. Сравнение проведено на разных наборах экспериментальных данных. В 2017 году начал работу новый лазер на свободных электронах European XFEL, который позволит регистрировать до 27000 дифракционных изображений в секунду. В статье представлены результаты исследования возможности применения разных подходов для классификации изображений в экспериментах на European XFEL в потоковом режиме.

Ключевые слова: *когерентная рентгеновская дифракционная микроскопия, биологические частицы, корреляционные коэффициенты, метод опорных векторов, метод к-средних, метод спектральной кластеризации, многослойный перцептрон, свёрточная нейронная сеть.*

1. ВВЕДЕНИЕ

Изучение трехмерной структуры белков и вирусов стало более эффективным с появлением метода когерентной рентгеновской дифракционной микроскопии (Coherent X-ray Diffractive Imaging – CXDI) [1, 2]. В экспериментах CXDI используется когерентное рентгеновское излучение лазеров на свободных электронах. С помощью этого метода можно определять структуру с разрешением до 1 Å [3, 4, 5, 6]. Знание структуры с таким разрешением позволит изучать механизмы функционирования исследуемых биологических объектов и может привести к новым открытиям в структурной биологии.

Рентгеновские импульсы лазеров на свободных электронах, используемые в

*s.bobkov@grid.kiae.ru

экспериментах CXDI, обладают уникальными характеристиками [7, 8, 9]. Пиковая интенсивность на 8 порядков превосходит пиковую интенсивность современных синхротронов. Длина волны рентгеновского импульса может достигать 1 Å, при этом длительность может быть менее 10 фемтосекунд.

В эксперименте идентичные экземпляры исследуемых объектов вводятся в луч лазера. Происходит упругое рассеяние на электронной плотности, рассеянное излучение регистрируется детектором на расстоянии около 1 метра от точки взаимодействия рентгеновского импульса и исследуемого объекта. Благодаря высокой интенсивности импульса, электроны выбиваются из атомов, и объект уничтожается в результате кулоновского взрыва. Несмотря на разрушение объекта, атомы не успевают значимо изменить свое положение за время взаимодействия, и поэтому получаемые дифракционные изображения соответствуют исходной структуре объекта [3, 4, 10, 11, 12].

В отличие от других методов структурной биологии (например, крио-электронной микроскопии), экземпляры изучаемых объектов вводятся в луч лазера в естественном состоянии, и появляется возможность изучения различных состояний биологических объектов.

Трехмерная структура восстанавливается на основе обработки множества дифракционных изображений исследуемого объекта в случайных ориентациях [13, 14, 15, 16]. Для достижения разрешения в 100 Å требуется собрать и обработать несколько тысяч изображений. Для улучшения разрешения до 10 Å необходимо на порядок больше изображений.

Не все получаемые изображения подходят для восстановления структуры. Многие из них пустые – ни одна частица не попала в луч лазера в момент импульса. В других случаях рентгеновский импульс рассеивается сразу на нескольких частицах. В случае рассеяния на одиночной частице, она может относиться как к экземпляру исследуемого объекта, так и к примеси или другим частицам. Для восстановления структуры подходят только изображения одиночных объектов изучаемого типа.

Фильтрация пустых изображений эффективно выполняется на основе анализа суммарной интенсивности сигнала на детекторе. В центре детектора сделан зазор для прохождения прямого пучка, если рентгеновский импульс не рассеивается на частице, он проходит через зазор и не попадает на чувствительную область детектора. Этот метод показывает хорошие результаты и успешно применяется в экспериментах.

Для фильтрации изображений примеси и изображений, содержащих дифракционную картину нескольких частиц, было предложено использовать классификацию по типу структуры [17]. Недавно, в работах [18, 19] была показана высокая эффективность для этой задачи метода классификации дифракционных изображений на основе корреляционных коэффициентов и метода опорных векторов.

В данной статье представлены результаты сравнительного анализа различных подходов на основе методов машинного обучения и нейронных сетей к задаче классификации по типу структуры на разных наборах экспериментальных данных. На основе полученных результатов делается заключение о возможности использования различных подходов для классификации по типу структуры в режиме потоковой обработки данных в экспериментах European XFEL.

2. ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ НАБОРОВ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

В работе использовались четыре набора дифракционных изображений, которые получены из открытой базы данных CXIDB [20]. В этой базе данных содержатся результаты экспериментов на лазерах на свободных электронах, которые предоставлены

авторами для свободного доступа. Вместе с дифракционными изображениями, авторами предоставлена информация о типе частиц на изображениях. Используемые в статье изображения были получены в экспериментах на установке LCLS (Стэнфорд) [7], на станции CAMP [21] с использованием детектора pnCCD [21]. В нашей работе на основе наборов данных из CXIDB были сформированы новые наборы, включающие изображения объектов нескольких типов. В этих наборах изображения случайно перемешаны, при этом для каждого изображения сохранена информация о типе частицы.

1. Набор CXIDB 13-14

Набор состоит из 958 дифракционных изображений. Среди них 590 изображений относятся к эксперименту CXIDB № 13 [22], в котором изучались образцы бактериофага T4 (*Escherichia virus* T4). Остальные 368 изображений относятся к эксперименту CXIDB № 14 [22], в котором исследовались образцы вируса *Paramecium bursarium chlorella virus* (PBCV) [23]. Оба типа биологических объектов имеют близкий размер, от 200 до 300 нм, поэтому классификация на основе анализа размера частиц невозможна.

2. Набор CXIDB 10-11

Набор состоит из 2149 дифракционных изображений, которые включают 1237 изображений нанориса (эллипсоид оксида железа под пленкой диоксида кремния), полученных в эксперименте CXIDB № 10 [22], и 912 изображения образцов магнетосом (магнитные частицы, которые встречаются в структуре бактерий), полученных в эксперименте CXIDB № 11 [22]. Образцы нанориса имеют вытянутую форму и размер 50 на 200 нм, размер образцов магнетосом составляет около 100 нм.

3. Набор CXIDB 20-25-37

Набор состоит из 2665 дифракционных изображений от трех типов частиц. 635 изображений относятся к результатам эксперимента CXIDB № 20 [24] и содержат изображения кластеров из двух слипшихся сфер полистирола (органический полимер). 1031 изображение относится к результатам эксперимента CXIDB № 25 [25] и содержит изображения очищенных карбоксисом (многогранные однослойные белковые тела полиэдрической формы, которые встречаются в структуре цианобактерий). Последние 999 изображений относятся к результатам эксперимента CXIDB № 37 [26] и содержат дифракционные изображения клеток цианобактерий *Cyanobium gracile*. В данном наборе объединены изображения, которые получены в разных экспериментах, разными группами ученых. Все изображения были дополнительно обработаны (Приложение 5), чтобы устранить специфические особенности разных экспериментов.

4. Набор CXIDB 25

Набор состоит из 4506 дифракционных изображений, полученных в эксперименте CXIDB № 25 [25] и содержит изображения очищенных карбоксисом (многогранные однослойные белковые тела полиэдрической формы, которые встречаются в структуре цианобактерий). В ходе предварительной обработки, набор был разделен на три группы: 2523 изображений одиночных экземпляров карбоксисом, 1477 изображений примесных частиц и 506 изображений, содержащих дифракционную картину нескольких экземпляров карбоксисом.

3. ТОЧНОСТЬ И ПОЛНОТА КЛАССИФИКАЦИИ

Для многих методов классификации сначала требуется провести обучение. В этих случаях наборы изображений разбиваются на две части: обучающий набор и проверочный

набор.

При проверке результатов классификации определяются следующие величины:
N_{tp} – количество изображений, которые относятся к данному типу и корректно классифицированы к данному типу;
N_{fp} – количество изображений, которые классифицированы к данному типу, однако относятся к другому типу;
N_{fn} – количество изображений, которые относятся к данному типу, но классифицированы к другим типам.

На их основании вычисляются критерии точности P [27] и полноты R [27] классификации:

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}}; \tag{1a}$$

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}}. \tag{1b}$$

Точность характеризует долю корректно классифицированных изображений и влияет на получаемое разрешение при восстановлении структуры. Полнота характеризует эффективность классификации и влияет на количество получаемых изображений.

Результат классификации подвергается перекрестной проверке по 10 блокам: полный набор изображений разделяется на 10 равных частей, затем 9 частей используются для обучения, а десятая часть используется для проверки результатов классификации. Сеансы обучения и проверки повторяются 10 раз, используя разные части для проверки. На статистике из 10 сеансов обучения и проверки определяются средние значения для точности и полноты классификации, а также их стандартные отклонения.

4. ПОДХОДЫ К КЛАССИФИКАЦИИ

Мы формулируем различные подходы к классификации, которые построены на основе следующих стандартных методов классификации с учителем: метод опорных векторов [28], классификация с помощью многослойного перцептрона [29] и классификация с помощью свёрточной нейронной сети [30].

Классификация изображений является сложной задачей из-за большой размерности, например, изображения детектора pnCCD состоят из 1048576 пикселей. Детектор AGIPD [31] на European XFEL имеет такое же разрешение. Поэтому важной составляющей процесса классификации является сжатие, оно позволяет уменьшить размерность данных на несколько порядков и тем самым ускорить обучение и классификацию. Для классификации удобно сжимать изображение в характеристический вектор, который сохраняет достаточную информацию о структуре.

В работе использовался метод сжатия на основе корреляционных коэффициентов, разработанный в статье [18] и подробно исследованный в статье [19]. Этот метод основан на теоретических результатах исследования процесса дифракции. Он выделяет особенности изображений, которые связаны со структурой исходных частиц и не связаны с ориентацией частиц на изображении и интенсивностью рентгеновского импульса. Метод сжатия был доработан для учета конструктивных особенностей экспериментальной установки: добавлен поиск центра симметрии дифракционных изображений (Приложение 3), а также разработан алгоритм расчета корреляционных коэффициентов с учетом зазоров детектора (Приложение 4). Длина характеристического вектора после сжатия составляет порядка 150 координат. Таким образом, размерность сокращается на 4 порядка.

Метод сжатия использовался совместно с методом опорных векторов для

классификации. Метод опорных векторов не позволяет достичь высокой точности без сжатия. Методы, использующие нейронные сети, показывают высокую точность без сжатия, поэтому они применялись напрямую к изображениям. При классификации изображений без сжатия, подход остается универсальным, его результаты более интересны, так как не используются дополнительные знания о процессе дифракции. Использование сжатия совместно с нейронными сетями ускоряет обучение и классификацию, но точность и полнота классификации остаются на прежнем уровне.

В литературе известны и другие методы классификации с учителем, например линейный [32] и квадратичный дискриминантный анализ [33]. Исследование этих методов показало, что они не позволяют проводить классификацию изображений по типу структуры без сжатия в характеристический вектор, а при использовании сжатия проигрывают методу опорных векторов.

Также были сформулированы подходы на основе метода к-средних [34, 35] и метода спектральной кластеризации [36], которые не требуют обучения. Помимо них были исследованы: метод Уорда [37], BIRCH [38], метод Mean Shift [39], метод распространения близости [40] и метод DBSCAN [41]. Точность и полнота классификации методом Уорда совпадает с результатами метода спектральной кластеризации на двух наборах данных, но на двух других наборах метод Уорда проигрывает по точности 10 % и более. Метод Mean Shift проигрывает методу спектральной кластеризации от двух до одиннадцати процентов по точности в зависимости от набора изображений. Методы DBSCAN и BIRCH проигрывают методу спектральной кластеризации от двух до семи процентов по точности, а при классификации набора CXIDB 20-25-37 оба метода не смогли определить разбиение на три кластера и показали точность классификации ниже 50 %. Метод распространения близости не дает удовлетворительных результатов классификации. По вышеизложенным причинам, упомянутые выше методы далее не рассматриваются.

1. Классификация на основе метода опорных векторов

В данном подходе метод опорных векторов используется для классификации характеристических векторов изображений, которые получены методом сжатия на основе корреляционных коэффициентов [18]. Метод опорных векторов для каждого элемента показывает вероятность, что тип элемента верно определен. Это значение мы будем называть «вероятностью корректной классификации». Точность подхода на основе метода опорных векторов может быть повышена за счет отбрасывания векторов, для которых вероятность корректной классификации ниже некоторого порогового значения. В статье представлены результаты двух подходов к классификации на основе метода опорных векторов: с порогом вероятности корректной классификации в 75 % и без порога. Значение 75 % выбрано для примера, оно может быть увеличено для большей точности или уменьшено для большей полноты классификации.

2. Классификация на основе метода к-средних

Данный подход использует метод к-средних для классификации характеристических векторов. Метод к-средних относится к группе методов кластеризации и находит оптимальное разбиение множества векторов на заданное число кластеров. Метод к-средних не использует обучающий набор для нахождения разбиения. Использование обучающего набора необходимо, чтобы сопоставить каждому кластеру наиболее подходящий класс.

3. Классификация на основе метода спектральной кластеризации

В данном подходе используется метод спектральной кластеризации [36] для классификации характеристических векторов. Метод спектральной кластеризации

находит спектр собственных значений и собственных векторов матрицы близости, на основе которых проводится кластеризация. Полученным кластерам сопоставляется наиболее подходящий класс с использованием обучающего набора.

4. Классификация на основе перцептрона с тремя скрытыми слоями

В данном подходе для классификации дифракционных изображений применяется искусственная нейронная сеть (ИНС) - перцептрон с тремя скрытыми слоями [29]. ИНС применяется к изображениям без сжатия, чтобы сохранить универсальность подхода.

Архитектура связей ИНС представлена на рисунке 1. Для активации нейронов скрытых слоев использовалась функция ReLU [42]. Для классификации на выходном слое использовалась функция Softmax [43].

На вход сети подавались дифракционные изображения с уменьшенным разрешением: группы 4×4 пикселя объединялись в один пиксель, сигнал которого равен сумме исходных сигналов. Таким образом, размерность уменьшалась в 16 раз. Обучение ИНС проводилось методом стохастического градиентного спуска. Параметры обучения указаны в таблице 1. После каждого скрытого слоя использовался метод регуляризации Dropout [44] с параметром 0.5, что позволяет избежать переобучения.

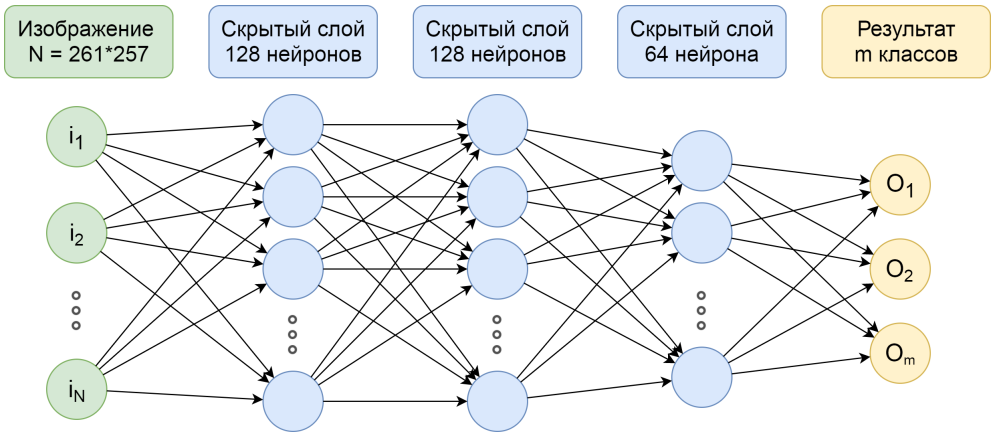


Рис. 1. Архитектура связей перцептрона с тремя скрытыми слоями.

Таблица 1. Параметры обучения нейронных сетей

Параметр	Значение
Скорость обучения	0.003
Размер пакета (batch)	200 изображений
Количество эпох обучения перцептрона с тремя скрытыми слоями	500
Количество эпох обучения свёрточной нейронной сети	300

5. Классификация на основе свёрточной нейронной сети

В данном подходе используются свёрточные нейронные сети для классификации дифракционных изображений. Свёрточные ИНС получили широкое распространение в задачах обработки изображений и машинного зрения, поэтому результаты применения свёрточных ИНС к дифракционным изображениям представляют большой интерес. Сверточные нейронные сети разработаны на основе моделирования работы зрительной коры головного мозга [30]. Такие ИНС используют специальные слои, где к изображениям

применяется операция свёртки с фиксированным ядром, что позволяет выделять особенности изображений без привязки к координатам пикселей. Традиционно, после слоя свёртки выполняется операция пулинга, уменьшающая размерность выходных данных, в нашем случае в 4 раза.

Архитектура связей свёрточной ИНС представлена на рисунке 2. Схема включает две группы, состоящие из слоя свёртки с последующим слоем пулинга. Размер ядра свёртки составляет 5 на 5 пикселей. Для каждого слоя свёртки использовалось 16 различных ядер. Каждый слой пулинга преобразует группы 2×2 пикселя в один пиксель следующего слоя, присваивая ему максимальный сигнал. Параметры обучения ИНС указаны в таблице 1. После каждого слоя пулинга использовался метод регуляризации Dropout [44] с параметром 0.25, что позволяет избежать переобучения.

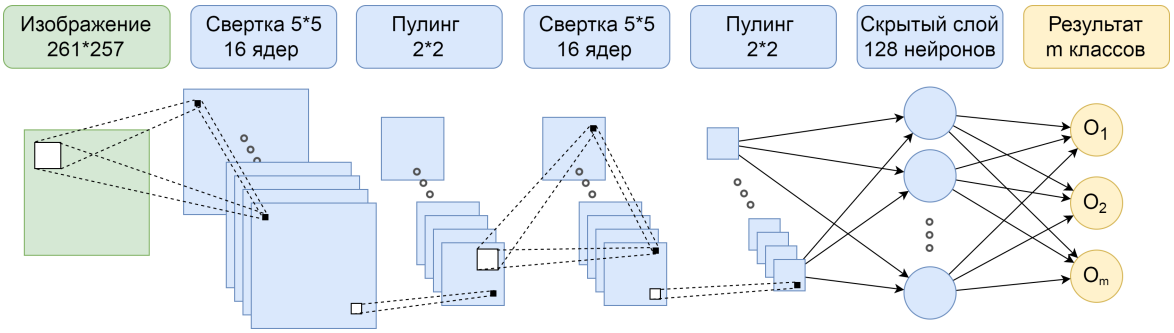


Рис. 2. Схема используемой свёрточной нейронной сети.

5. РЕЗУЛЬТАТЫ АНАЛИЗА РАЗЛИЧНЫХ ПОДХОДОВ К КЛАССИФИКАЦИИ

1. Набор CXIDB 13-14

Результаты классификации 958 изображений набора CXIDB 13-14 на два типа представлены в таблице 2. Заметим, что наилучшие результаты классификации двух типов достигаются при использовании подхода на основе метода опорных векторов, он показывает значения средней точности и полноты более 90 %. При использовании порога вероятности корректной классификации в 75 %, точность возрастает на 3 % при уменьшении полноты на 10 %.

Таблица 2. Результаты классификации набора CXIDB 13-14

Подходы к классификации	Тип частицы	Точность	Полнота
Классификация на основе метода опорных векторов с порогом 75 %	бактериофаг T4	94.4 % ± 3.6 %	86.1 % ± 3.8 %
	PBCV	93.5 % ± 4.4 %	71.3 % ± 9.2 %
Классификация на основе метода опорных векторов без порога	бактериофаг T4	91.6 % ± 4.2 %	94.0 % ± 3.8 %
	PBCV	89.7 % ± 7.0 %	85.7 % ± 7.5 %
Классификация на основе метода к-средних	бактериофаг T4	74.2 % ± 3.3 %	99.0 % ± 1.1 %
	PBCV	96.7 % ± 3.6 %	44.9 % ± 5.1%
Классификация на основе метода спектральной кластеризации	бактериофаг T4	79.6 % ± 5.0 %	98. 3% ± 1.5 %
	PBCV	95.7 % ± 3.8 %	59.5 % ± 9.5 %
Классификация на основе перцептрона с тремя скрытыми слоями	бактериофаг T4	86.2 % ± 3.5 %	93.5 % ± 2.8 %
	PBCV	88.4 % ± 5.2 %	76.6 % ± 6.7 %
Классификация на основе свёрточной нейронной сети	бактериофаг T4	87.2 % ± 2.2 %	94.7 % ± 3.2 %
	PBCV	90.5 % ± 6.1 %	78.3 % ± 4.8 %

2. Набор CXIDB 10-11

Результаты классификации набора CXIDB 10-11 на два типа представлены в таблице 3. Наилучшие результаты классификации достигаются в подходе на основе метода опорных векторов, точность и полнота классификации превышают 99.6 %. Использовании порога вероятности корректной классификации не влияет на результаты.

Таблица 3. Результаты классификации набора CXIDB 10-11

Подходы к классификации	Тип частицы	Точность	Полнота
Классификация на основе метода опорных векторов с порогом 75 %	нанорис	100.0 % ± 0.0 %	99.7 % ± 0.7 %
	магнетосомы	99.6 % ± 1.0 %	100.0 % ± 0.0 %
Классификация на основе метода опорных векторов без порога	нанорис	100.0 % ± 0.0 %	99.7 % ± 0.7 %
	магнетосомы	99.6 % ± 1.0 %	100.0 % ± 0.0 %
Классификация на основе метода к-средних	нанорис	83.8 % ± 4.3 %	34.7 % ± 2.9 %
	магнетосомы	54.1 % ± 2.8 %	79.2 % ± 12.8 %
Классификация на основе метода спектральной кластеризации	нанорис	97.8 % ± 1.8 %	27.4 % ± 2.5 %
	магнетосомы	53.8 % ± 1.9 %	88.0 % ± 12.8 %
Классификация на основе перцептрона с тремя скрытыми слоями	нанорис	97.1 % ± 1.4 %	98.8 % ± 0.8 %
	магнетосомы	98.4 % ± 1.0 %	95.9 % ± 2.0 %
Классификация на основе свёрточной нейронной сети	нанорис	99.2 % ± 0.9 %	99.6 % ± 0.4 %
	магнетосомы	99.5 % ± 0.5 %	98.8 % ± 1.4 %

3. Набор CXIDB 20-25-37

Результаты классификации набора CXIDB 20-25-37 на три типа представлены в таблице 4. Наилучшие результаты достигаются при использовании подхода на основе свёрточной нейронной сети, средние значения точности и полноты классификации составляют более 99 %. Точность и полнота классификации для подхода на основе метода опорных векторов ниже на 0.2 %. Использование порогового значения вероятности корректной классификации в 75 % повышает среднюю точность на 0.3 % и понижает среднюю полноту на 0.8 %.

4. Набор CXIDB 25

Результаты классификации набора CXIDB 25 представлены в таблице 5. В отличие от остальных методов, здесь ставится задача фильтрации изображений одиночных карбоксисом от изображений примеси и изображений, которых содержат вклад от нескольких частиц. Поэтому важна точность и полнота классификации одиночных карбоксисом, а не средняя точность и полнота.

Подход на основе метода опорных векторов показывает точность и полноту классификации в 90–91 %, а при использовании порога вероятности корректной классификации в 75 %, точность возрастает до 95 % при полноте в 79 %. Точность классификации одиночных карбоксисом в других подходах ниже.

5. Итоговые результаты

По результатам анализа различных подходов к классификации на разных наборах данных можно сделать следующие выводы. Все рассмотренные подходы показывают высокую точность классификации. Подход к классификации на основе метода опорных векторов, в общем, показывает наилучшие значения точности и полноты. Использование порога вероятности корректной классификации позволяет повысить точность на несколько процентов за счет уменьшения полноты.

Таблица 4. Результаты классификации набора CXIDB 20-25-37

Подходы к классификации	Тип частицы	Точность	Полнота
Классификация на основе метода опорных векторов с порогом 75 %	сферы полистирола	99.8 % ± 0.5 %	99.2 % ± 0.8 %
	карбоксисомы	100.0 % ± 0.0 %	98.3 % ± 1.1 %
	Cyanobium gracile	100.0 % ± 0.0 %	100.0 % ± 0.0
Классификация на основе метода опорных векторов без порога	сферы полистирола	98.3 % ± 0.8 %	100.0 % ± 0.0 %
	карбоксисомы	100.0 % ± 0.0 %	98.9 % ± 0.6 %
	Cyanobium gracile	100.0 % ± 0.0 %	100.0 % ± 0.0
Классификация на основе метода к-средних	сферы полистирола	94.6 % ± 3.4 %	85.9 % ± 4.2 %
	карбоксисомы	67.3 % ± 3.0 %	99.7 % ± 0.6 %
	Cyanobium gracile	100.0% ± 0.0%	56.2% ± 3.3
Классификация на основе метода спектральной кластеризации	сферы полистирола	98.9 % ± 1.6 %	80.8 % ± 5.2 %
	карбоксисомы	59.7 % ± 2.6 %	99.9 % ± 0.3 %
	Cyanobium gracile	100.0 % ± 0.0 %	42.1 % ± 2.6
Классификация на основе перцептрона с тремя скрытыми слоями	сферы полистирола	98.8 % ± 1.9 %	100.0 % ± 0.0 %
	карбоксисомы	99.7 % ± 0.7 %	100.0 % ± 0.0 %
	Cyanobium gracile	100.0 % ± 0.0 %	98.9 % ± 1.4
Классификация на основе свёрточной нейронной сети	сферы полистирола	99.7 % ± 1.0 %	100.0 % ± 0.0
	карбоксисомы	99.8 % ± 0.4 %	99.9 % ± 0.3 %
	Cyanobium gracile	99.9 % ± 0.3 %	99.6 % ± 0.7

Отметим, что подход на основе свёрточной нейронной сети показывает сравнимую точность и полноту классификации. Подход на основе перцептрона с тремя скрытыми слоями проигрывает три-четыре процента по точности и полноте классификации.

Подходы к классификации на основе метода к-средних и метода спектральной кластеризации заметно проигрывают либо в точности, либо в полноте классификации, но они не требуют обучения.

6. ОПТИМИЗАЦИЯ РАЗМЕРА ОБУЧАЮЩЕГО НАБОРА

Размер обучающего набора определяет временные затраты на разметку изображений и обучение. Такие затраты важны, например при классификации в режиме потоковой обработки. Поэтому минимизация размера обучающего набора представляет практический интерес. В этом разделе исследовалась зависимость точности классификации от размера обучающего набора.

В представленных выше (раздел 5) результатах классификации, обучающий набор составлял 90 % от полного набора (раздел 3). Далее в статье термин «максимальная точность» соответствует точности при размере обучающего набора в 90 % от полного набора. Мы приводим результаты исследования точности классификации для меньших размеров обучающего набора, чтобы оценить возможность уменьшить его размер, сократив временные затраты. Полное время обработки изображений складывается из трех этапов:

– разметка изображений обучающего набора, которая проводится вручную,
– обучение,
– классификация.

На рисунке 3 представлена зависимость точности классификации от размера обучающего набора для разных наборов изображений. Для каждого значения размера обучающего набора было проведено 20 сеансов обучения и проверки результатов. На каждом сеансе случайно определялся обучающий набор заданного размера, остальные изображения использовались для проверки. На статистике из 20 сеансов определялась

Таблица 5. Результаты классификации для набора CXIDB 25

Подходы к классификации	Тип частицы	Точность	Полнота
Классификация на основе метода опорных векторов с порогом 75 %	карбоксисомы	95.0 % ± 0.7 %	79.6 % ± 3.9 %
	остальные	96.8 % ± 0.9 %	70.7 % ± 3.5 %
Классификация на основе метода опорных векторов без порога	карбоксисомы	90.7 % ± 1.3 %	91.5 % ± 1.8 %
	остальные	89.1 % ± 1.9 %	88.0 % ± 2.4 %
Классификация на основе метода к-средних	карбоксисомы	64.4 % ± 2.9 %	99.8 % ± 0.4 %
	остальные	98.9 % ± 1.8 %	29.8 % ± 4.0 %
Классификация на основе метода спектральной кластеризации	карбоксисомы	68.9 % ± 2.2 %	99.3 % ± 0.3 %
	остальные	98.0 % ± 1.0 %	42.9 % ± 3.6 %
Классификация на основе перцептрона с тремя скрытыми слоями	карбоксисомы	86.6 % ± 1.7 %	91.2 % ± 2.1 %
	остальные	88.0 % ± 2.3 %	82.1 % ± 2.4 %
Классификация на основе свёрточной нейронной сети	карбоксисомы	85.8 % ± 2.3 %	95.5 % ± 1.2 %
	остальные	93.4 % ± 1.3 %	79.7 % ± 3.6 %

средняя точность классификации и стандартное отклонение точности. Полученная зависимость точности от размера обучающего набора аппроксимировались степенной функцией с помощью метода наименьших квадратов.

Подходы на основе метода к-средних и метода спектральной кластеризации не используют информацию о типе изображений в обучающем наборе, поэтому их точность не зависит от размера обучающего набора. В остальных подходах к классификации точность растет с увеличением размера обучающего набора.

Определим оптимальный размер обучающего набора, при котором средняя точность классификации составляет 99 % от максимальной точности (получаемой при обучении на 90%-й части от полного набора). В таблице 6 приведены оптимальные размеры обучающего набора для разных подходов и наборов изображений.

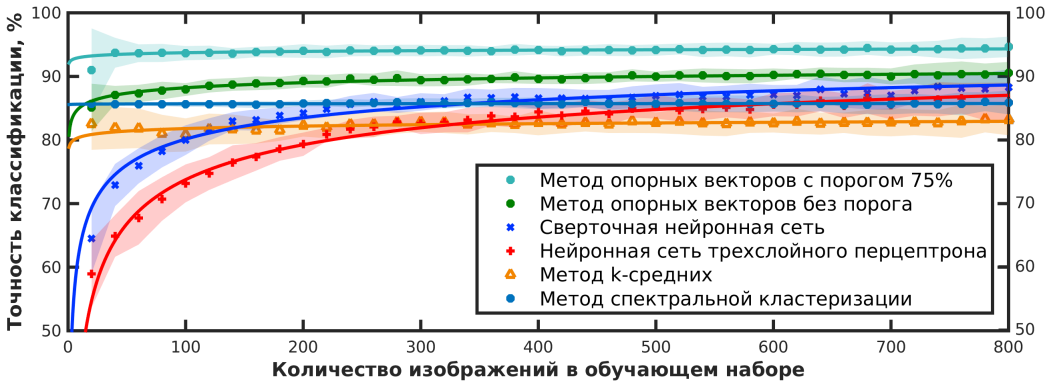
Можно заметить, что для подхода на основе метода опорных векторов, использование порога вероятности корректной классификации в разы сокращает оптимальный размер обучающего набора. В сравнении с методами основанными на нейронных сетях, оптимальный размер обучающего набора меньше в 10 и более раз.

Таблица 6. Оптимальный размер обучающего набора на разных наборах данных

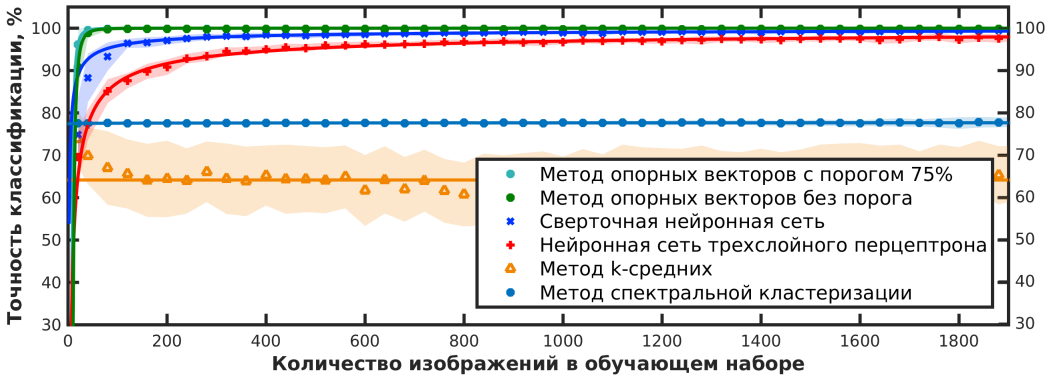
Подходы к классификации	CXIDB 13-14	CXIDB 10-11	CXIDB 20-25-37	CXIDB 25
Классификация на основе метода опорных векторов с порогом 75 %	40	30	160	80
Классификация на основе метода опорных векторов без порога	300	40	600	300
Классификация на основе перцептрона с тремя скрытыми слоями	640	920	520	2720
Классификация на основе свёрточной нейронной сети	580	420	520	2680
Всего изображений в наборе	958	2149	2665	4506

7. ВРЕМЕННЫЕ ЗАТРАТЫ НА ОБУЧЕНИЕ И КЛАССИФИКАЦИЮ

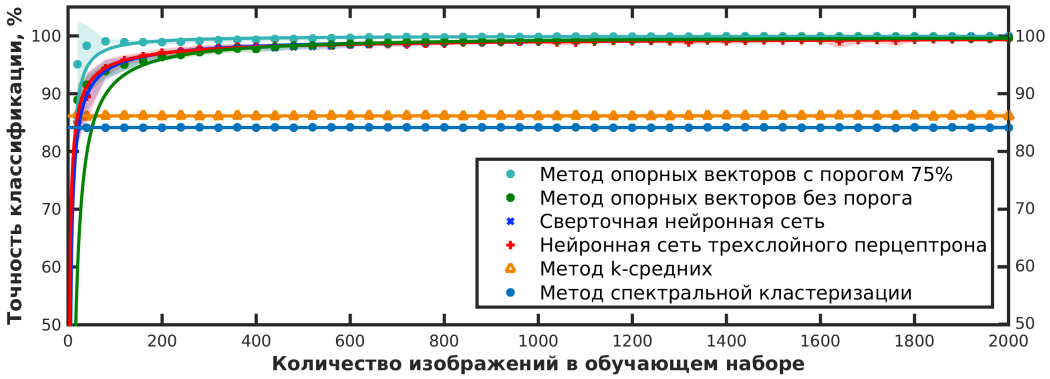
Временные затраты на обучение и классификацию зависят от используемых методов и аппаратных ресурсов. В наших исследованиях мы использовали сервера с центральным



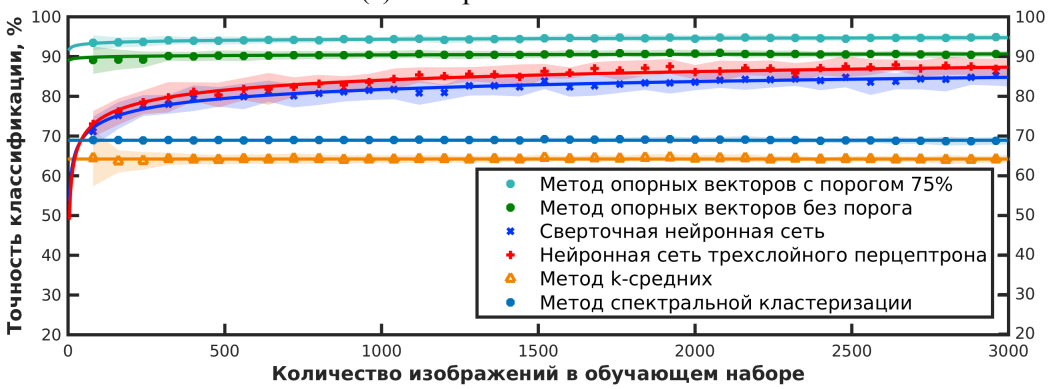
(a) Набор CXIDB 13-14



(b) Набор CXIDB 10-11



(c) Набор CXIDB 20-25-37



(d) Набор CXIDB 25

Рис. 3. Зависимость точности классификации от размера обучающего набора на разных наборах изображений. Закрашенная область соответствует стандартному отклонению $(-\sigma, +\sigma)$.

процессором Intel Xeon E5-2680v3 и графическими процессорами NVIDIA Tesla K80. Временные затраты для рассматриваемых подходов к классификации были измерены и усреднены по всем наборам данных. Результаты приведены в таблице 7. Время обучения определялось для оптимальных размеров обучающих наборов, приведенных в таблице 6.

Время классификации 1000 изображений для подходов на основе метода опорных векторов, метода к-средних и метода спектральной кластеризации, которые используют сжатие изображений в характеристический вектор, составляет 51 секунду. При использовании графических процессоров, это время сокращается в 4 раза. Для нейронных сетей, время классификации составляет около 20 секунд, а использование графических процессоров ускоряет классификацию незначительно.

Время обучения для подходов к классификации на основе сжатия в характеристический вектор составляет секунды, при переходе на графические процессоры, оно сокращается в 4.5 раза. В тоже время, для подхода на основе перцептрона с тремя скрытыми слоями время обучения составляет несколько минут, а для подхода на основе свёрточной нейронной сети – несколько часов.

Заметим, что на этапах классификации и сжатия присутствует естественный параллелизм, разные изображения можно обрабатывать параллельно, используя многопоточность на нескольких узлах и на нескольких процессорах внутри узлов.

Таким образом, все рассмотренные подходы позволяют классифицировать 1000 изображений за секунды, ни один подход не имеет решающего преимущества. При обучении появляется серьезная разница, подходы с использованием метода сжатия тратят на обучение секунды, а подходы на основе нейронных сетей без сжатия – минуты и часы. Кроме того, использование многопоточности естественно приведет к ускорению классификации и сжатия.

Таблица 7. Временные затраты на обучение и классификацию

Подходы к классификации	Время классификации 1000 изображений		Время обучения, секунды	
	CPU	GPU	CPU	GPU
Классификация на основе метода опорных векторов с порогом 75 %	51.3	11.3	4.5	1.0
Классификация на основе метода опорных векторов без порога	51.3	11.3	15.9	3.5
Классификация на основе метода к-средних	51.2	11.2	2.3	0.51
Классификация на основе метода спектральной кластеризации	51.9	11.7	2.3	0.50
Классификация на основе перцептрона с тремя скрытыми слоями	18.5	17.2	353	287
Классификация на основе свёрточной нейронной сети	23.9	19.7	8021	1542

8. ВОЗМОЖНЫЙ СЦЕНАРИЙ ПОТОКОВОЙ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ В ЭКСПЕРИМЕНТАХ EUROPEAN XFEL

Лазер на свободных электронах European XFEL будет регистрировать до 27000 изображений в секунду. Можно ожидать, что около 1 % регистрируемых изображений будут содержать дифракционную картину, остальные изображения пустые, их можно быстро и надежно отфильтровать на основе анализа суммарной интенсивности. Таким образом, в экспериментах на European XFEL будет регистрироваться порядка 270

Таблица 8. Характеристики аппаратных ресурсов, достаточные для потоковой классификации дифракционных изображений в экспериментах European XFEL

Подходы к классификации	Классификация на центральном процессоре	Классификация на графическом процессоре
Классификация на основе метода опорных векторов, метода к-средних и метода спектральной кластеризации	Сервер (Intel Xeon) с 10 ядрами на частоте 2.5 ГГц	Сервер с двумя графическими процессорами NVIDIA K80
Классификация с использованием нейронных сетей	Сервер (Intel Xeon) с 6 ядрами на частоте 2.5 ГГц	Сервер с тремя графическими процессорами NVIDIA K80

Таблица 9. Оценка времени ручной разметки обучающего набора оптимального размера

Подходы к классификации	Время работы для разметки обучающего набора			
	CXIDB 13-14	CXIDB 10-11	CXIDB 20-25-37	CXIDB 25
Классификация на основе метода опорных векторов с порогом 75 %	3 мин	3 мин	12 мин	7 мин
Классификация на основе метода опорных векторов без порога	25 мин	3 мин	50 мин	25 мин
Классификация на основе перцептрона с тремя скрытыми слоями	50 мин	1 час	45 мин	4 часа
Классификация на основе свёрточной нейронной сети	50 мин	35 мин	45 мин	4 часа

дифракционных изображений в секунду (без учета пустых изображений), которые можно классифицировать по типу структуры. Объем одного изображения составляет 2 Мбайт для детектора AGIPD [31]. Поэтому основное требование к ресурсам для потоковой обработки – возможность приема данных на скорости в 0.6 Гбайт/сек. За 12 часов эксперимента будет регистрироваться 23 Тбайт данных без учета пустых изображений.

На основе анализа временных затрат, представленных в разделе 7, мы можем определить достаточные характеристики аппаратных ресурсов, которые представлены в таблице 8. Характеристики были рассчитаны теоретически, а затем проверены в вычислительном эксперименте.

Однако в таком эксперименте мы не учитываем время, требуемое для ручной разметки обучающего набора. Можно оценить время разметки на уровне 5 секунд на одно изображение при работе эксперта. Тогда, время ручной разметки обучающих наборов оптимального размера представлено в таблице 9.

Таким образом, потоковая классификация может быть реализована по следующему сценарию:

1. После начала работы, накапливается обучающий набор оптимального размера (таблица 6). Характеристические вектора изображений рассчитываются в режиме онлайн.

2. Ручная разметка обучающего набора. Временные затраты приведены в таблице 9.

3. Обучение. Временные затраты приведены в таблице 7.

4. После завершения обучения, новые дифракционные изображения классифицируются в режиме онлайн. Сохраненные ранее изображения могут быть классифицированы впоследствии.

Заметим, что в таком сценарии могут использоваться все рассмотренные подходы, однако оптимально использовать подход на основе метода опорных векторов, так как он показывает высокую точность, а оптимальный обучающий набор имеет минимальный размер. Стоит отметить, что в результате накопления системных ошибок, этап разметки и обучения, возможно, нужно будет повторять.

9. ВЫВОДЫ

Подход к классификации на основе метода опорных векторов стабильно показывает наилучшую скорость, точность и полноту классификации. Использование порога вероятности корректной классификации позволяет повысить точность еще на 3–5 %. Высокая точность и полнота классификации обусловлена выбранным методом сжатия изображений в характеристический вектор, который уменьшает размерность данных на четыре порядка и сохраняет информацию, связанную с структурой исходных частиц.

Использование метода к-средних и метода спектральной кластеризации позволяет пропустить этап разметки обучающего набора.

Подходы на основе нейронных сетей позволяют достичь высокой точности и полноты классификации без сжатия в характеристический вектор, который учитывает особенности процесса дифракции. Точность и полнота классификации в подходах на основе нейронных сетей лишь немного уступает подходу на основе метода опорных векторов, однако подходы на основе нейронных сетей без сжатия требуют разметки обучающего набора большого размера.

Все рассмотренные подходы могут использоваться для потоковой классификации изображений в экспериментах на European XFEL, но оптимально использовать подход на основе метода опорных векторов.

Работа была выполнена при поддержке соглашения о предоставлении субсидии № 14.616.21.0003 (уникальный идентификатор научных исследований RFMEFI61614X0003). Работа была выполнена с использованием оборудования центра коллективного пользования «Комплекс моделирования и обработки данных исследовательских установок мега-класса» НИЦ «Курчатовский институт», http://ckp.nrcki.ru. Автор выражает благодарность Ильину В. А., Вартаньянцу И. А., Сбоеву А. Г., Рыбке Р. Б. и Теслюку А. Б. за помощь в работе и в подготовке статьи.

ПРИЛОЖЕНИЕ 1.

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ ФИЛЬТРАЦИИ ИЗОБРАЖЕНИЙ НЕСКОЛЬКИХ ЧАСТИЦ

Хотя изображения нескольких частиц не подходят для восстановления структуры в экспериментах CXDI, такие изображения позволяют определить параметры структуры изучаемых частиц на основе метода XCCA [45, 46, 47, 48].

Возможность фильтрации изображений нескольких частиц от остальных изображений исследовалась на наборе CXIDB 25. Классификация проводилась на два типа: изображения, содержащие дифракционный вклад нескольких частиц, и остальные, к ним относятся изображения одиночных карбоксисом и примесных частиц. Результаты представлены в таблице 10.

Таблица 10. Результаты классификации изображений нескольких частиц для набора CXIDB 25

Подход	Тип частицы	Точность	Полнота
Классификация на основе метода опорных векторов с порогом 75 %	несколько частиц	81.7 % ± 11.4 %	20.0 % ± 5.6 %
	остальные	92.6 % ± 1.2 %	97.4 % ± 0.8 %
Классификация на основе метода опорных векторов без порога	несколько частиц	78.6 % ± 9.5 %	24.4 % ± 5.4 %
	остальные	91.2 % ± 1.2 %	99.2 % ± 0.4 %
Классификация на основе перцептрона с тремя скрытыми слоями	несколько частиц	62.9 % ± 8.4 %	18.0 % ± 9.7 %
	остальные	90.7 % ± 0.9 %	98.2 % ± 0.7 %
Классификация на основе свёрточной нейронной сети	несколько частиц	65.9 % ± 7.7 %	34.6 % ± 9.5 %
	остальные	92.3 % ± 0.9 %	97.5 % ± 1.0 %

Подход на основе метода опорных векторов, показывает точность классификации на уровне 79 % при полноте 25 %. Использование порога вероятности корректной классификации в 75 % позволяет повысить точность до 81% при уменьшении полноты до 20 %. Такой подход можно использовать для классификации, однако низкая полнота делает его применение неэффективным.

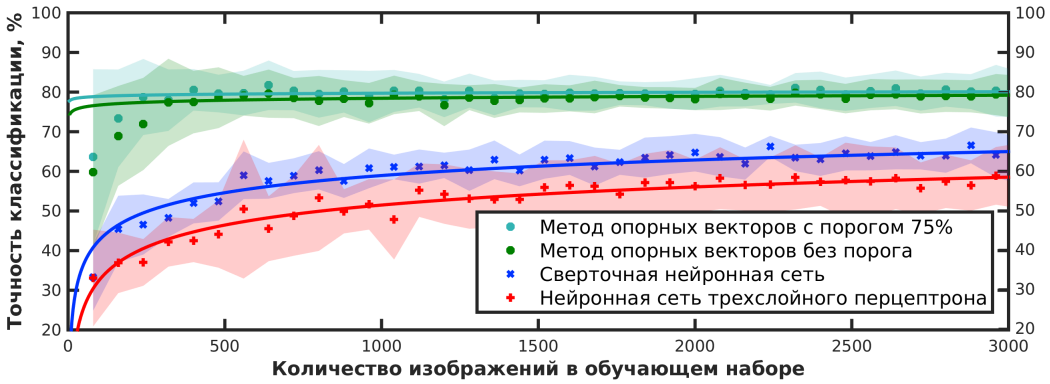


Рис. 4. Зависимость точности фильтрации изображений нескольких частиц от размера обучающего набора для набора CXIDB 25. Закрашенная область соответствует стандартному отклонению $(-\sigma, +\sigma)$.

При классификации на основе нейронных сетей, точность не превышает 66 %, однако полнота классификации составляет 35 %. Низкая точность ограничивает применимость

таких подходов. Остальные подходы показывают результаты хуже, поэтому они не подходят для решения задачи фильтрации изображений нескольких частиц.

Для эффективной фильтрации изображений нескольких частиц требуется разработка нового подхода.

Зависимость точности от размера обучающего набора представлена на рисунке 4. Подход к классификации на основе метода опорных векторов достигает точности в 99 % от максимума при размере обучающего набора в 950 изображений. При использовании порога вероятности корректной классификации в 75 %, оптимальный размер обучающего набора составляет 210 изображений. Для походов на основе нейронных сетей, оптимальный размер обучающего набора превышает 3000 изображений.

ПРИЛОЖЕНИЕ 2.
ПОИСК ЦЕНТРА СИММЕТРИИ ДИФРАКЦИОННОЙ КАРТИНЫ

Амплитуда рассеяния излучения в экспериментах CXDI может быть описана в рамках первого приближения Борна:

$$A(\boldsymbol{q}) = A_0 \int \rho(\boldsymbol{r}) e^{i\boldsymbol{q}\boldsymbol{r}} d\boldsymbol{r}, \tag{2}$$

где момент передачи импульса $\boldsymbol{q} = \boldsymbol{k}_{out} - \boldsymbol{k}_{in}$ определяется разницей волновых векторов падающего и рассеянного излучения. Из формулы 2 следует, что $|A(\boldsymbol{q})| = |A(-\boldsymbol{q})|$, то есть модуль амплитуды и интенсивность рассеяния обладает центральной симметрией относительно вектора передачи импульса. Сечение пространства значений вектора $\boldsymbol{q}$ дифракционным изображением можно можно считать плоскостью для малых значений $\boldsymbol{q}$. Так как интенсивность дифракционного изображения падает как третья степень модуля $\boldsymbol{q}$, будем считать, отклонением сечения от плоскости можно пренебречь, и все дифракционные изображения обладают центральной симметрией.

Положение центра симметрии дифракционной картины относительно детектора связано с проекцией положения частицы в момент рассеяния на плоскость детектора. Анализ дифракционных изображений, полученных в эксперименте, показывает, что центр симметрии меняет свое положение для разных изображений, смещение центра симметрии относительно детектора может составлять до 10 пикселей. Такое смещение вносит заметный вклад в корреляционные коэффициенты, и, следовательно, в получаемые характеристические векторы.

Наилучшие результаты определения центра симметрии дифракционной картины относительно детектора были достигнуты в следующем подходе: определяется такое положение центра симметрии, при котором инверсия относительно центра симметрии меняет дифракционное изображение наименьшим образом:

1. $I(x, y)$ - интенсивность дифракционного изображения. x и y – координаты пикселей на детекторе.

2. Обозначим координаты центра симметрии дифракционной картины относительно детектора δ_x и δ_y. Тогда интенсивность инвертированного изображения равна:

$$\bar{I}(x, y, \delta_x, \delta_y) = I(-x + 2\delta_x, -y + 2\delta_y). \tag{3}$$

   Так как x и y принимают целые значения, δ_x, δ_y имеют вид $\frac{n}{2}$, где n – целое.

3. Для заданных значений δ_x и δ_y определяется функция разницы прямого и инвертированного изображения:

$$D(\delta_x, \delta_y) = \frac{\langle I(x, y) - \bar{I}(x, y, \delta_x, \delta_y)\rangle_{x,y}}{\langle I(x, y) + \bar{I}(x, y, \delta_x, \delta_y)\rangle_{x,y}}. \tag{4}$$

   Важно отметить, что при расчете $D(\delta_x, \delta_y)$ используются только такие значения x и y, при которых и I и $\bar{I}$ не попадают в зазор детектора.

4. Пункты 1–3 повторяются для разных значений δ_x и δ_y, пока не будет найден минимум функции $D(\delta_x, \delta_y)$. Минимум соответствует наиболее вероятному положению центра симметрии.

Использование данного метода поиска центра симметрии позволило увеличить точность

классификации, так как характеристические векторы при сжатии лучше соответствуют структуре исходных объектов.

ПРИЛОЖЕНИЕ 3.

РАСЧЕТ КОРРЕЛЯЦИОННЫХ КОЭФФИЦИЕНТОВ С УЧЕТОМ ЗАЗОРОВ ДЕТЕКТОРА

При расчете характеристических векторов определяются корреляционные коэффициенты для дифракционного изображения в полярных координатах:

$$C(q,\Delta) = \langle I(q,\varphi)\cdot I(q,\varphi+\Delta)\rangle_{\varphi}\,, \tag{5}$$

где q и φ – радиус и угол полярной системы координат, а Δ – угол корреляции. Наличие зазоров в дифракционной картине оказывает заметное влияние на значения корреляционных коэффициентов, особенно для малых углов рассеяния, где интенсивность приближается к максимуму рабочего диапазона детектора. Если угол Δ равен углу между областями зазоров в полярных координатах, функция $C(q,\Delta)$ получает дополнительный вклад, превосходящий вклад от структуры, который представляет основной интерес. Для используемых сегодня детекторов, зазоры в дифракционной картине вызывают появление пиков функции $C(q,\Delta)$ для Δ кратных π. Такие пики приводят к изменениям во множестве компонент Фурье разложения и в характеристических векторах изображений.

Для уменьшения влияния зазоров детектора на значения корреляционных коэффициентов существуют несколько подходов, они сравнивались на наборе CXIDB 13-14 и следующий подход показал наилучшую точность: интенсивность внутри зазора приравнивалась средней интенсивности вне зазора для заданного радиуса q в полярных координатах. В этом подходе расчет корреляционных коэффициентов состоит из следующих этапов:

1. Перевод дифракционного изображения в полярные координаты
2. Для каждого значения радиуса q определяется средняя интенсивность вне зазора:

$$I_{gap}(q) = \langle I(q,\varphi)\rangle_{\varphi}\,,\ \varphi \notin \text{зазор}. \tag{6}$$

   Интенсивность внутри зазора приравнивается $I_{gap}(q)$ для данного q.
3. Проводится расчет корреляционных коэффициентов и характеристического вектора, считая что зазоры теперь отсутствуют.

Данный подход позволяет увеличить точность классификации дифракционных изображений, так как характеристические векторы лучше соответствуют структуре исходных объектов для дифракционных изображений.

ПРИЛОЖЕНИЕ 4.

ПОДГОТОВКА ДИФРАКЦИОННЫХ ИЗОБРАЖЕНИЙ ИЗ НАБОРА CXIDB 20-25-37

Входящие в набор CXIDB 20-25-37 дифракционные изображения были получены разными группами ученых в разное время. На таких изображениях присутствуют особенности, связанные с отличиями в используемых параметрах эксперимента и в предварительной обработке данных. Также на детекторе присутствовали дефекты, картина которых отличалась для разных экспериментов. Такие различия позволяют определить тип образца на изображении с точностью около 100 %, не прибегая к анализу дифракционной картины. Однако при классификации изображений одного эксперимента, все дополнительные метки, облегчающие классификацию, отсутствуют.

Все изображения третьего набора были предварительно обработаны единым образом для устранения различий. Таким образом, результаты классификации позволят оценить применимость рассматриваемых подходов к классификации поступающих в течение эксперимента изображений согласно типу исходной структуры.

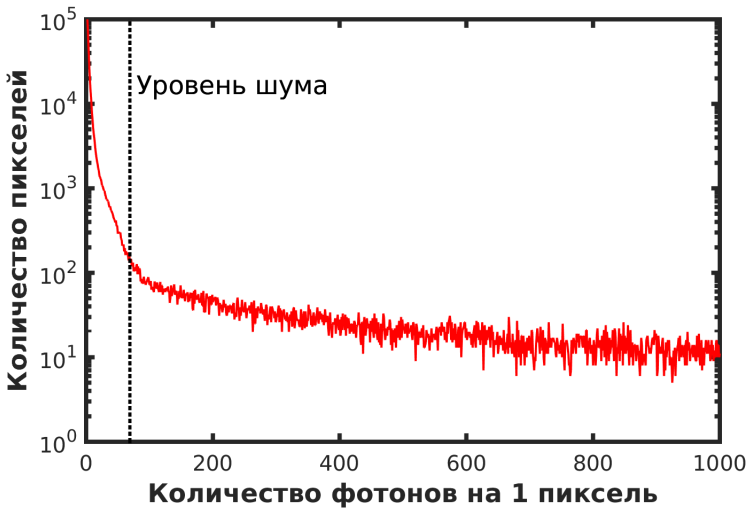


Рис. 5. Пример гистограммы интенсивности дифракционного изображения.

Предварительная обработка состояла из двух этапов:

1. Уменьшение уровня шума на изображениях.

2. Наложение маски дефектов.

На первом этапе предварительной обработки убирались различия в уровне шума, связанные с работой детектора. Такой шум присутствует на всех изображениях и его распределение интенсивности равномерно для всех пикселей. Изображения, полученные в разных экспериментах, отличаются по уровню шума, так как использовались разные параметры калибровки детектора. Для каждого изображения строилась гистограмма интенсивности 5. На гистограмме наблюдается две области: область низкой интенсивности, ниже уровня шума и область высокой интенсивности, где полезный сигнал превалирует над шумом. В области низкой интенсивности гистограмма имеет резкий пик, схожий с распределением Пуассона, в области высокой интенсивности, напротив, количество пикселей для каждого значения интенсивности меняется не значительно, максимальная разница не превышает 10 для всей области. Граничный уровень шума устанавливался равным интенсивности, при которой значение гистограммы

равно 0.001 от максимального значения. Полученное значение уровня шума вычиталось из интенсивности каждого пикселя изображения, при отрицательном результате, интенсивность устанавливалась равной нулю.

На втором этапе убирались различия в дефектах детектора. Для этого определялись области, где изображения имели высокую интенсивность независимо от дифракционной картины. Такие дефекты также связаны с работой детектора, но из-за высокой интенсивности не убираются на первом шаге. Была создана маска, которая учитывала дефекты для всех трех используемых наборов CXIDB, она представлена на рисунке 6. Также в маску включена область зазора в центре детектора, так как ширина зазора немного отличалась для разных изображений, и круглая область в центре, где детектор регистрирует край лазерного луча, картина которого также отличалась. Для всех изображений интенсивность под маской была установлена равной нулю.

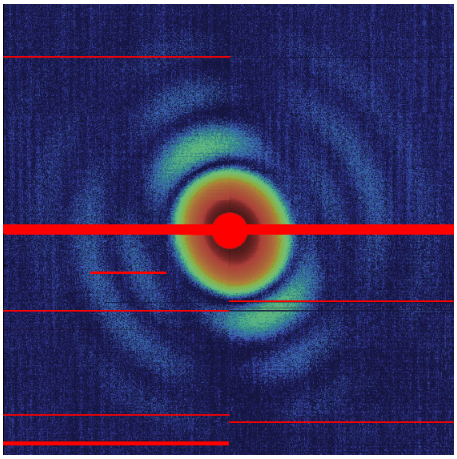


Рис. 6. Используемая маска дефектов детектора для предварительной обработки изображений.

СПИСОК ЛИТЕРАТУРЫ

1. Miao J., Ishikawa T., Johnson B., Anderson E.H., Lai B., Hodgson K.O. High resolution 3D x-ray diffraction microscopy. *Physical Review Letters*. 2002. V. 89. № 8. P. 088303. doi: 10.1103/physrevlett.89.088303
2. Chapman H.N., Nugent K.A. Coherent lensless X-ray imaging. *Nature Photonics*. 2010. V. 4. № 12. P. 833–839.
3. Chapman H.N., Barty A., Bogan M.J., Boutet S., Frank M., Hau-Riege S.P., Marchesini S., Woods B.W., Bajt S., Benner W.H. et al. Femtosecond diffractive imaging with a soft-X-ray free-electron laser. *arXiv preprint physics/0610044*. 2006. doi: 10.1038/nphoton.2010.240
4. Gaffney K.J., Chapman H.N. Imaging atomic structure and dynamics with ultrafast X-ray scattering. *Science*. 2007. V. 316. № 5830. P. 1444–1448. doi: 10.1126/science.1135923
5. Seibert M.M., Ekeberg T., Maia F.R., Svenda M., Andreasson J., Jönsson O., Odić D., Iwan B., Rocker A., Westphal D. et al. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*. 2011. V. 470. № 7332. P. 78–81. doi: 10.1038/nature09748
6. Mancuso A.P., Yefanov O.M., Vartanyants I.A. Coherent diffractive imaging of biological samples at synchrotron and free electron laser facilities. *Journal of Biotechnology*. 2010. V. 149. № 4. P. 229–237. doi: 10.1016/j.jbiotec.2010.01.024
7. Emma P., Akre R., Arthur J., Bionta R., Bostedt C., Bozek J., Brachmann A., Bucksbaum P., Coffee R., Decker F. et al. First lasing and operation of an ångstrom-wavelength free-electron laser. *Nature Photonics*. 2010. V. 4. № 9. P. 641–647. doi: 10.1038/nphoton.2010.176

8. Ishikawa T., Aoyagi H., Asaka T., Asano Y., Azumi N., Bizen T., Ego H., Fukami K., Fukui T., Furukawa Y. et al. A compact X-ray free-electron laser emitting in the sub-angstrom region. *Nature Photonics*. 2012. V. 6. № 8. P. 540–544. doi: 10.1038/nphoton.2012.141
9. *The European X-Ray Free-Electron laser*: Technical Design Report. Eds. Massimo A. et al. Hamburg, Germany: European XFEL project team, 2007.
10. Neutze R., Wouts R., van der Spoel D., Weckert E., Hajdu J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*. 2000. V. 406. № 6797. P. 752–757. doi: 10.1038/35021099
11. Lorenz U., Kabachnik N.M., Weckert E., Vartanyants I.A. Impact of ultrafast electronic damage in single-particle x-ray imaging experiments. *Physical Review E*. 2012. V. 86. № 5. P. 051911. doi: 10.1103/physreve.86.051911
12. Gorobtsov O.Y., Lorenz U., Kabachnik N.M., Vartanyants I.A. Theoretical study of electronic damage in single-particle imaging experiments at x-ray free-electron lasers for pulse durations from 0.1 to 10 fs. *Physical Review E*. 2015. V. 91. № 6. P. 062712. doi: 10.1103/physreve.91.062712
13. Ne-Te Duane Loh, Veit Elser Reconstruction algorithm for single-particle diffraction imaging experiments. *Physical Review E*. 2009. V. 80. № 2. doi: 10.1103/physreve.80.026705
14. Fienup J.R. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*. 1978. V. 3. № 1. P. 27–29. doi: 10.1364/ol.3.000027
15. Fienup J.R. Phase retrieval algorithms: a comparison. *Appl. Opt.* 1982. V. 21. № 15. P. 2758. doi: 10.1364/ao.21.002758
16. Chen C., Miao J., Wang C., Lee T. Application of optimization technique to noncrystalline x-ray diffraction microscopy: Guided hybrid input-output method. *Physical Review B*. 2007. V. 76. № 6. P. 064113. doi: 10.1103/physrevb.76.064113
17. Yoon C.H., Schwander P., Abergel C., Andersson I., Andreasson J., Aquila A., Bajt S., Barthelmess M., Barty A., Bogan M.J. et al. Unsupervised classification of single-particle X-ray diffraction snapshots by spectral clustering. *Optics Express*. 2011. V. 19. № 17. P. 16542–16549. doi: 10.1364/OE.19.016542
18. Bobkov S.A., Teslyuk A.B., Kurta R.P., Gorobtsov O.Y., Yefanov O.M., Ilyin V.A., Senin R.A., Vartanyants I.A. Sorting algorithms for single-particle imaging experiments at X-ray free-electron lasers. *Journal of Synchrotron Radiation*. 2015. V. 22. P. 1345–1352. doi: 10.1107/S1600577515017348
19. Бобков С.А., Теслюк А.Б., Вартаньянц И.А., Ильин В.А. Классификация дифракционных изображений биологических макромолекул с разными типами симметрии в экспериментах по когерентной рентгеновской дифракционной микроскопии. *Математическая биология и биоинформатика*. 2016. Т. 11. № 2. С. 299–310. doi: 10.17537/2016.11.299
20. Maia F.R.N.C. The Coherent X-ray Imaging Data Bank. *Nature Methods*. 2012. V. 9. № 9. P. 854–855. doi: 10.1038/nmeth.2110
21. Strüder L., Epp S., Rolles D., Hartmann R., Holl P., Lutz G., Soltau H., Eckart R., Reich C., Heinzinger K. et al. Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. 2010. V. 614. № 3. P. 483–496. doi: 10.1016/j.nima.2009.12.053
22. Kassemeyer S., Steinbrener J., Lomb L., Hartmann E., Aquila A., Barty A., Martin A.V., Hampton C.Y., Bajt S., Barthelmess M. et al. Femtosecond free-electron laser x-ray diffraction data sets for algorithm development. *Optics Express*. 2012. V. 20. № 4.

P. 4149–4158. doi: 10.1364/OE.20.004149
23. Van Etten J.L., Burbank D.E., Xia Y., Meints R.H. Growth cycle of a virus, PBCV-1, that infects Chlorella-like algae. *Virology*. 1983. V. 126. № 1. P. 117–125. doi: 10.1016/0042-6822(83)90466-x
24. Starodub D., Aquila A., Bajt S., Barthelmess M., Barty A., Bostedt C., Bozek J.D., Coppola N., Doak R.B., Epp S.W. et al. Single-particle structure determination by correlations of snapshot X-ray diffraction patterns. *Nature Communications*. 2012. V. 3. P. 1276. doi: 10.1038/ncomms2288
25. Hantke M.F., Hasse D., Maia F.R., Ekeberg T., John K., Svenda M., Loh N.D., Martin A.V., Timneanu N., Larsson D.S. et al. High-throughput imaging of heterogeneous cell organelles with an X-ray laser. *Nature Photonics*. 2014. V. 8. № 12. P. 943–949. doi: 10.1038/nphoton.2014.270
26. Van Der Schot G., Svenda M., Maia F.R., Hantke M.F., DePonte D.P., Seibert M.M., Aquila A., Schulz J., Kirian R.A., Liang M. et al. Open data set of live cyanobacterial cells imaged using an X-ray laser. *Scientific Data*. 2016. V. 3. doi: 10.1038/sdata.2016.58
27. Ting K.M. Precision and Recall. In: *Encyclopedia of Machine Learning*. Springer, 2010. P. 781. ISBN 978-0-387-30164-8.
28. Cortes C., Vapnik V. Support-vector networks. *Machine Learning*. 1995. V. 20. № 3. P. 273–297.
29. Rosenblatt F. *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. 1961. doi: 10.21236/ad0256582
30. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. V. 521. № 7553. P. 436–444. doi: 10.1038/nature14539
31. Henrich B., Becker J., Dinapoli R., Goettlicher P., Graafsma H., Hirsemann H., Klanner R., Krueger H., Mazzocco R., Mozzanica A. et al. The adaptive gain integrating pixel detector AGIPD a detector for the European XFEL. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. 2011. V. 633. P. S11–S14. doi: 10.1016/j.nima.2010.06.107
32. Fisher R.A. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*. 1936. V. 7. № 2. P. 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x
33. Cover T.M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*. 1965. № 3. P. 326–334. doi: 10.1109/pgec.1965.264137
34. Steinhaus H. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*. 1956. V. 1. № 804. P. 801.
35. Lloyd S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*. 1982. V. 28. № 2. P. 129–137. doi: 10.1109/tit.1982.1056489
36. Shi J., Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2000. V. 22. № 8. P. 888–905. doi: 10.1109/34.868688
37. Ward Jr J.H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*. 1963. V. 58. № 301. P. 236–244. doi: 10.2307/2282967
38. Zhang T., Ramakrishnan R., Livny M. BIRCH: an efficient data clustering method for very large databases. In: *Proceeding SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data.* 1996. V. 25. № 2. P. 103–114. doi: 10.1145/233269.233324
39. Cheng Y. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1995. V. 17. № 8. P. 790–799. doi: 10.1109/34.400568
40. Frey B.J., Dueck D. Clustering by passing messages between data points. *Science*. 2007. V. 315. № 5814. P. 972–976. doi: 10.1126/science.1136800

41. Ester M., Kriegel H., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *KDD-96 Proceedings*. 1996. V. 96. № 34. P. 226–231.
42. Hahnloser R.H.R., Sarpeshkar R., Mahowald M.A., Douglas R.J., Seung H.S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*. 2000. V. 405. № 6789. P. 947–951.
43. Bishop C.M. *Pattern recognition and machine learning*. 2006.
44. Hinton G.E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*. 2012.
45. Altarelli M., Kurta R.P., Vartanyants I.A. X-ray cross-correlation analysis and local symmetries of disordered systems: General theory. *Physical Review B*. 2010. V. 82. № 10. P. 104207. doi: 10.1103/PhysRevB.82.104207
46. Kurta R.P., Dronyak R., Altarelli M., Weckert E., Vartanyants I.A. Solution of the phase problem for coherent scattering from a disordered system of identical particles. *New Journal of Physics*. 2013. V. 15. № 1. P. 013059.
47. Pedrini B., Menzel A., Guizar-Sicairos M., Guzenko V., Gorelick S., David C., Patterson B.D., Abela R. Two-dimensional structure from random multiparticle X-ray scattering images using cross-correlations. *Nature Communications*. 2013. V. 4. P. 1647. doi: 10.1038/ncomms2622
48. Saldin D.K., Poon H., Schwander P., Uddin M., Schmidt M. Reconstructing an icosahedral virus from single-particle diffraction experiments. *Optics Express*. 2011. V. 19. № 18. P. 17318–17335. doi: 10.1364/oe.19.017318

Рукопись поступила в редакцию 01.11.2017
Дата опубликования 29.11.2017